

Constant conditional entropy and related hypotheses

Ramon Ferrer-i-Cancho¹, Łukasz Dębowski² and Fermín Moscoso del Prado Martín³

¹ Complexity & Quantitative Linguistics Lab
 Departament de Llenguatges i Sistemes Informàtics,
 TALP Research Center, Universitat Politècnica de Catalunya,
 Campus Nord, Edifici Omega Jordi Girona Salgado 1-3.
 08034 Barcelona, Catalonia (Spain)

² Institute of Computer Science
 Polish Academy of Sciences
 ul. Jana Kazimierza 5
 01-248 Warszawa, Poland

³ Department of Linguistics
 University of California at Santa Barbara
 Santa Barbara, CA 93106-3100, USA

E-mail: rferrericancho@lsi.upc.edu, ldebowsk@ipipan.waw.pl and
 fermin.moscoso-del-prado@gmail.com

Abstract. Constant entropy rate (conditional entropies must remain constant as the sequence length increases) and uniform information density (conditional probabilities must remain constant as the sequence length increases) are two information theoretic principles that are argued to underlie a wide range of linguistic phenomena. Here we revise the predictions of these principles to the light of Hilberg’s law on the scaling of conditional entropy in language and related laws. We show that constant entropy rate (CER) and two interpretations for uniform information density (UID), full UID and strong UID, are inconsistent with these laws. Strong UID implies CER but the reverse is not true. Full UID, a particular case of UID, leads to costly uncorrelated sequences that are totally unrealistic. We conclude that CER and its particular cases are incomplete hypotheses about the scaling of conditional entropies.

Keywords: constant entropy rate, uniform information density, Hilberg’s law.

PACS numbers: 89.70.-a Information and communication theory
 89.75.Da Systems obeying scaling laws
 05.40.-a Fluctuation phenomena, random processes, noise, and Brownian motion
 02.30.Lt Sequences, series, and summability

1. Introduction

Uniform information density and constant entropy rate [1, 2] are two information theoretic principles that have been put forward to explain various linguistic phenomena, e.g., syntactic reduction [1, 3] and the frequency of word orders [4]. In order to present these principles, we provide some definitions. Formally, u is a linguistic sequence of n elements, e.g., words or letters, i.e. $u = x_1, \dots, x_n$, where x_i is the i -th elements of the sequence. U_i is defined as the set of elements that can appear in the i -th position of the sequence and U as the set of all possible sequences of length n , i.e. $U \subseteq U_1 \times \dots \times U_n$. Both U and all the U_i 's are support sets, i.e. all their members have non-zero probability and all non-members have zero probability. X_i is the random variable that takes values from the set of elements U_i . Hereafter, x_i stands for an element of U_i or a value of X_i by default.

The constant entropy rate (CER) hypothesis states that the conditional entropy of an element given the previous elements remains constant [2]. To express it formally, $H(X_i|X_1, \dots, X_{i-1})$ is defined as the Shannon conditional entropy of the i -th element of a sequence given the $i - 1$ preceding elements. The CER states that $H(X_i|X_1, \dots, X_{i-1})$ remains constant as i increases ($i = 1, 2, \dots, n$), i.e.

$$H(X_1) = H(X_2|X_1) = \dots = H(X_n|X_1, \dots, X_{n-1}). \quad (1)$$

The uniform information density (UID) hypothesis [1, 3] states that the conditional probability of an element given the previous elements should remain constant. To express it formally, $p(x_i|x_1, \dots, x_{i-1})$ is defined as the probability of x_i given the preceding elements in u . We say than a particular utterance $u = x_1, \dots, x_n$ satisfies the UID condition if and only if

$$p(x_1) = p(x_2|x_1) = \dots = p(x_n|x_1, \dots, x_{n-1}). \quad (2)$$

Here we will review the validity of CER and UID to the light of the real scaling of the conditional and other entropies as n increases, or equivalently, as i , the length of the prefix, increases [5, 6, 7, 8, 9]. We will show that the scaling of these entropies is inconsistent with CER and two interpretations of UID, strong UID and full UID, a particular case of strong UID. In essence, our arguments are the following. First, CER is inconsistent with Hilberg's law, a hypothesis on the entropy of natural language made on the base of Shannon's famous experiment [5]. Hilberg's law states that $H(X_n|X_1, \dots, X_{n-1}) \sim n^{-\beta}$, with $\beta \approx 1/2$ [6], whereas CER means $\beta = 0$. Second, strong UID, i.e., all the utterances in U satisfy the UID condition, is a particular case of CER. The latter can be easily shown. By taking logarithms and multiplying by $p(u) = p(x_1, \dots, x_n)$ on the UID condition (Eq. 2), the strong UID can be written equivalently as

$$\begin{aligned} p(u) \log p(x_1) &= p(u) \log p(x_2|x_1) = \dots \\ &= p(u) \log p(x_n|x_1, \dots, x_{n-1}) \end{aligned}$$

for any $u = x_1, \dots, x_n \in U$. Summing over all $u \in U$ (the utterances that do not belong to U have no effect thanks to the convention $0 \log 0 = 0$ [10]) and inverting the sign, the strong UID leads to

$$\begin{aligned} - \sum_u p(u) \log p(x_1) &= - \sum_u p(u) \log p(x_2|x_1) = \dots \\ &= - \sum_u p(u) \log p(x_n|x_1, \dots, x_{n-1}), \end{aligned}$$

where each part corresponds to the definition of a Shannon (conditional) entropy, i.e. the definition of CER in Eq. 1 is recovered.

These ideas are developed in the coming sections, which will not only examine the meaning of strong UID but also that of full UID, a particularly degenerated version of real language where sequences of symbols are uncorrelated and entropies are maximum.

2. The uniform information density hypothesis

First, let us inspect some consequences of the UID hypothesis. One of the major challenges of the uniform density hypothesis is defining a criterion for the applicability of the hypothesis. UID was originally defined on a single sequence [1]. The utility and power of the hypothesis depends on its scope: the more sequences UID concerns, the better. We start with a very ambitious UID hypothesis, namely that UID holds for any sequence in U that can be formed combining the elements of U_i , i.e.

$$U = U_1 \times \dots \times U_n. \quad (3)$$

We call it full UID. We also consider a weaker but still strong version, where UID holds also for any sequence in U but Eq. 3 does not need to be satisfied. This version is called strong UID. In fact, full UID implies strong UID but the reverse is not true. To see the latter, consider support set $U = \{(a, b), (a, c), (d, e), (d, f)\}$, where $p(X_2 = x_2|X_1 = x_1) = p(X_1 = x_1) = 1/2$ for any $(x_1, x_2) \in U$. We thus have strong UID but the full UID fails to hold since $U \neq U_1 \times U_2 = \{a, d\} \times \{b, c, e, f\}$.

The UID condition can be written in terms of the joint probability. By the chain rule of conditional probability

$$p_u = p(x_1)p(x_2|x_1)\dots p(x_n|x_1, \dots, x_{n-1}). \quad (4)$$

and Eq. 2 we obtain that the UID hypothesis implies

$$p(x_1, \dots, x_n) = p(x_1)^n. \quad (5)$$

That is, strong UID means that all sequences beginning with the same word are equally likely. Furthermore, noting that, by definition, we have

$$p(x_i) = \sum_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n} p(x_1, \dots, x_n) \quad (6)$$

and applying Eq. 5, we obtain

$$p(x_1) = p(x_1)^n \sum_{\substack{x_2, \dots, x_n, \\ (x_1, \dots, x_n) \in U}} 1 = p(x_1)^n |U(X_1 = x_1)|$$

where $U(X_1 = x_1)$ is the subset of U containing all the sequences where $X_1 = x_1$. Therefore,

$$p(x_1) = |U(X_1 = x_1)|^{\frac{1}{1-n}}. \quad (7)$$

Taking in particular $n = 2$, we obtain that $p(x_1)$ must be a reciprocal of a natural number.

2.1. Full UID

With the help of the properties of the UID hypothesis above it is easy to show that full UID, implies, that for any sequence x_1, \dots, x_n ,

(i) x_1, \dots, x_n are independent, i.e.,

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i).$$

(ii) The sets of elements that can appear at each position of the sequence have the same cardinality, i.e. $|U_1| = \dots = |U_n|$.

(iii) All words occurring in the same position are equally likely, i.e. $p(x_i) = 1/|U_1|$.

If UID holds for any sequence beginning with x_1 that can be formed by combining elements from U_2, \dots, U_n , then $|U(X_1 = x_1)| = \prod_{i=2}^n |U_i|$ and Eq. 7 becomes

$$p(x_1) = \left[\prod_{i=2}^n |U_i| \right]^{\frac{1}{1-n}}. \quad (8)$$

for $n \geq 2$. Eq. 8 indicates that $p(x_1)$ is the same for any $x_1 \in U_1$. Thus, the condition

$$\sum_{x_1 \in X_1} p(x_1) = 1$$

gives $p(x_1) = 1/|U_1|$ and the UID condition in Eq. 5 becomes

$$p(x_1, \dots, x_n) = |U_1|^{-n}. \quad (9)$$

Now we will derive $p(x_i)$ for $i \geq 2$. Employing Eq. 9, Eq. 6 can be written as (assuming $i \geq 2$)

$$p(x_i) = |U_1|^{1-n} \frac{\prod_{j=2}^n |U_j|}{|U_i|} = |U_1|^{1-n} \prod_{j=2}^{i-1} |U_j| \prod_{j=i+1}^n |U_j|.$$

Notice that, again, $p(x_i)$ is the same for any $x_i \in X_i$. Thus, the condition

$$\sum_{x_i \in X_i} p(x_i) = 1$$

gives $p(x_i) = 1/|U_i|$ for any $i = 1, \dots, n$. Now, we will show that $|U_i| = |U_1|$ for any $i = 1, \dots, n$. By definition, we have

$$p(x_1, \dots, x_{i-1}) = \sum_{x_i} p(x_1, \dots, x_i),$$

which, thanks to UID (recall Eq. 9), becomes

$$p(x_1, \dots, x_{i-1}) = |U_1|^{-i} \sum_{x_i \in U_i} 1 = |U_1|^{-i} |U_i|.$$

Applying UID again (Eq. 9) to the l.h.s, gives $|U_1|^{1-i} = |U_1|^{-i} |U_i|$, and thus $|U_1| = |U_i|$ and $p(x_i) = 1/|U_1|$ for any $i = 1, \dots, n$. Now it is easy to show that x_1, \dots, x_n are independent. The definition of independence, i.e. $p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i)$, holds trivially under UID since then $p(x_1, \dots, x_n) = |U_i|^n$ and $p(x_i) = 1/|U_i|$ for any $i = 1, \dots, n$. Finally, notice that full UID means that $p(X_i = x_i) = 1/|U_1|$ but does not imply that the elements of the sequences (regardless of their position) are equally likely. If all the sequences of $U = \{(a, a), (a, c), (b, a), (b, c)\}$ have probability $1/4$, one then has full UID but the probability of producing a is $4/8$ while the probability of producing c is $2/8$.

2.2. The relationship between CER and strong UID

Unravelling the relationship between CER and UID is also in need. For instance, [4] is based on the idea of UID but its mathematical implementation is based on the definition of CER. Strong UID implies CER (Section 1) but it will be shown that the reverse implication does not hold. Consider CER with $n = 2$, i.e. $H(X_2|X_1) = H(X_1)$, and independence between X_1 and X_2 , i.e. $H(X_2|X_1) = H(X_2)$. Thus, $H(X_1) = H(X_2)$. Assume also that $U_1 = \{a, b\}$ and $U_2 = \{c, d\}$. If one has $p(X_1 = a) = p(X_2 = c) = 2/5$, $p(X_1 = b) = p(X_2 = d) = 3/5$, then one has CER but strong UID does not hold because $2/5$ is not the reciprocal of a natural number, a condition for strong UID noticed beforehand.

3. The real scaling of entropies versus constant entropy rate

A serious consequence of the properties of full UID is that it is totally unrealistic with respect to natural language for several reasons. First, full UID leads to a sequence of independent elements, while long range correlations pervade linguistic sequences both at the level of letters and the level of words, e.g. [11, 7, 12, 13, 14]. Second, full UID is problematic because entropies are maximum. $H(X_i)$ is maximum for any $i = 1, \dots, n$ as $H(X_i) = \log |U_i|$. The joint entropy is also maximum because [10]

$$H(X_1, \dots, X_n) \leq \sum_{i=1}^n H(X_i) \quad (10)$$

in general but full UID transforms the inequality of Eq. 10 into a mere equality because the elements making a sequence are independent. Since entropy is a measure of cognitive cost [15, 16], full UID means the entropy related costs are maximum.

Last but not least, the plausibility of UID or CER for natural language is undermined by the results of celebrated experiments. Let $H_e(X_n|X_1, \dots, X_{n-1})$ be an estimate of $H(X_n|X_1, \dots, X_{n-1})$ from real data and ε_n the error of the estimate, i.e.,

$$H_e(X_n|X_1, \dots, X_{n-1}) = H(X_n|X_1, \dots, X_{n-1}) + \varepsilon_n,$$

where $\varepsilon_n \geq 0$ in general (by the nonnegativity of Kullback-Leibler divergence and Kraft inequality, it follows that the errors ε_n are positive if entropy is estimated by means of universal probability or universal coding [10]). Hilberg [6] reanalyzed Shannon's estimates of conditional entropy for English [5] and discovered that

$$H_e(X_n|X_1, \dots, X_{n-1}) \approx C_e n^{\alpha-1} \quad (11)$$

with $C_e > 0$, $\alpha \approx 0.5$, and $n \leq 100$ characters. Extrapolating Hilberg's law, Eq. 11, for $n \gg 100$ requires some caution. If accepted with $\varepsilon_n = 0$, Eq. 11 would imply asymptotic determinism of human utterances with an entropy rate $h = \lim_{n \rightarrow \infty} H(X_n|X_1, \dots, X_{n-1})$ equal to 0. Thus it is more plausible to accept that the scaling law of the true entropy (not its estimate) is

$$H(X_n|X_1, \dots, X_{n-1}) \approx C n^{\alpha-1} + h, \quad (12)$$

with $C < C_e$ and a sufficiently small constant $h > 0$. That the modified Hilberg law, Eq. 12, is indeed valid for natural language for $n \gg 100$ characters can be corroborated by the following fact: the modified Hilberg's law implies a lower bound for the growth of V , the observed vocabulary size, as a function of the T , the text length. Namely, Eq. 12 implies that V grows at least as $\sim T^\alpha / \log T$ [17], which is in good accordance with the real growth of V [18].

In contrast, CER is a competing hypothesis on $H(X_n|X_1, \dots, X_{n-1})$, the true conditional entropy. Assuming that CER, Eq. 1, holds in spite of Hilberg's law, Eq. 11, is equivalent to stating that the errors $\varepsilon_n \approx C_e n^{\alpha-1} - H(X_1)$ are systematically decreasing as n increases and negative for moderate n if $H(X_1)$ is large enough. This seems unrealistic since, as we have stated above, the errors of entropy estimates should be positive in general and it is unlikely that the errors are systematically diminishing (undersampling, a very important source of error, usually increases as n increases).

The disagreement between real language and CER concerning the decay of conditional entropy can be rephrased in terms of other entropic measures: $H(X_1, \dots, X_n)$ the joint or block entropy [8, 9] and $H(X_1, \dots, X_n)/n$, the joint entropy per unit [7]. It is easy to infer the scaling of these entropies from the modified Hilberg law, Eq. 12, by means of the chain rule of the joint entropy [10], which yields

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i|X_1, \dots, X_{i-1}) \approx \int_0^n [C m^{\alpha-1} + h] dm = \alpha^{-1} C n^\alpha + hn, \quad (13)$$

and thus

$$H(X_1, \dots, X_n)/n \approx \alpha^{-1} C n^{\alpha-1} + h. \quad (14)$$

In contrast, CER predicts a linear growth of the joint entropy, i.e.

$$H(X_1, \dots, X_n)/n = H(X_1), \quad (15)$$

which can be proven by applying the definition of CER in Eq. 1 to the chain rule of joint entropy [10].

As an independent confirmation of Shannon's research, according to Ref. [7], the estimates of $H(X_1, \dots, X_n)/n$ for sequences of letters from an English novel show good

agreement with Eq. 14 with $\alpha = 0.5$. Furthermore, the estimates of $H(X_1, \dots, X_n)$ grow sublinearly with n not only for texts in English but also for sequences from many other languages with different kinds of units [9, 8]. CER predicts that $H(X_1, \dots, X_n)/n$ is constant, Eq. 15, a strikingly different result. Therefore, the inconsistencies of CER are robust in the sense that do not depend on entropic measure, the language or the units of the sequence being considered. The same inconsistencies concern full and strong UID as it has been shown in Sections 1 and 2 that Full UID \Rightarrow Strong UID \Rightarrow CER.

4. Discussion

We have shown that CER and the two interpretations of the UID hypothesis (full UID and strong UID) are inconsistent with the scaling law for entropy of natural language called Hilberg’s law. Future research should address the challenge of what modifications of UID/CER can be consistent with real language. In order to save UID/CER, we envisage that probabilities and conditional entropies for real language stem from a conflict between principles: one acting towards UID/CER and another acting against UID/CER. A similar conflict between principles has been hypothesized for Zipf’s law for word frequencies, namely that the frequency of the i -th most frequent word of a text is $\sim i^{-\tau}$ [19]: the law emerges from a conflict between two principles, minimization of the entropy of words and maximization of the mutual information between words and meanings where none of the principles is totally realized [20, 21]. Indeed, Refs. [17, 22] show that there is a close relationship between Zipf’s law and Hilberg’s law so the conflict of principles that leads to Zipf’s law may be the same that prevents UID/CER from the full realization.

It is worth noting that there are simple information theoretic principles that lead to UID/CER. For instance, minimization of conditional entropy, leads to Eq. 1 with $H(X_1) = 0$. Interestingly, entropy minimization can be easily justified as it implies the minimization of cognitive cost [15, 16] and is used to explain Zipf’s law for word frequencies [20, 21]. Clearly, this conditional entropy minimization could not be acting alone as its total realization implies that all possible sequences have probability zero except one and therefore Hilberg’s law could not hold. As the force towards UID/CER is not working alone, the nature of the second factor in conflict must be clarified. The view of UID/CER as a principle in conflict means that all the currently available explanations of linguistic phenomena based upon UID/CER, e.g., [2, 1, 4], are *a priori* incomplete.

Acknowledgments

We are grateful to R. Levy, F. Jaeger, S. Piantadosi and E. Gibson for helpful discussions. This work was supported by the grant *Iniciació i reincorporació a la recerca* from the Universitat Politècnica de Catalunya and the grants BASMATI (TIN2011-27479-C04-03) and OpenMT-2 (TIN2009-14675-C03) from the Spanish Ministry of Science and Innovation (RFC).

References

- [1] R. Levy and T. F. Jaeger. Speakers optimize information density through syntactic reduction. *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, 2007.
- [2] D. Genzel and E. Charniak. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 199–206, 2002.
- [3] T. F. Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23 – 62, 2010.
- [4] L. Maurits, A. A. Perfors, and D. Navarro. Why are some word orders more common than others? a uniform information density account. In *Advances in Neural Information Processing Systems*, volume 23, pages 1585–1593, 2010.
- [5] C. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50–64, 1951.
- [6] W. Hilberg. Der bekannte Grenzwert der redundanzfreien Information in Texten – eine Fehlinterpretation der Shannonschen Experimente? *Frequenz*, 44:243–248, 1990.
- [7] W. Ebeling and T. Pöschel. Entropy and long-range correlations in literary English. *Europhysics Letters*, 26(4):241–246, 1994.
- [8] A. O. Schmitt and H. Herzel. Estimating the entropy of DNA sequences. *Journal of Theoretical Biology*, 188(3):369 – 377, 1997.
- [9] R. Rao. Probabilistic analysis of an ancient undeciphered script. *IEEE Computer*, 43:76–80, 2010.
- [10] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley, New York, 2006. 2nd edition.
- [11] M. Montemurro and P. A. Pury. Long-range fractal correlations in literary corpora. *Fractals*, 10:451–461, 2002.
- [12] E. Alvarez-Lacalle, B. Dorow, J.-P. Eckmann, and E. Moses. Hierarchical structures induce long-range dynamical correlations in written texts. *Proceedings of the National Academy of Sciences*, 103:7956–7961, 2006.
- [13] F. Moscoso del Prado Martín. The universal ‘shape’ of human languages: spectral analysis beyond speech. *PLoS ONE*, page in press, 2011.
- [14] E. A. Altmann, G. Cristadoro, and M. D. Esposti. On the origin of long-range correlations in texts. *Proc. Natl. Acad. Sci. USA*, 109:11582–11587, 2012.
- [15] F. Moscoso del Prado Martín, Alexandar Kostić, and R. H. Baayen. Putting the bits together: an information theoretical perspective on morphological processing. *Cognition*, 94:1–18, 2004.
- [16] R. Ferrer-i-Cancho and A. Díaz-Guilera. The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics*, page P06009, 2007.
- [17] Ł. Dębowski. On the vocabulary of grammar-based codes and the logical consistency of texts. *IEEE Transactions on Information Theory*, 57:4589–4599, 2011.
- [18] G. Herdan. *Quantitative linguistics*. Butterworths, London, 1964.
- [19] G. K. Zipf. *Human behaviour and the principle of least effort*. Addison-Wesley, Cambridge (MA), USA, 1949.
- [20] R. Ferrer i Cancho. Zipf’s law from a communicative phase transition. *European Physical Journal B*, 47:449–457, 2005.
- [21] M. Prokopenko, N. Ay, O. Obst, and D. Polani. Phase transitions in least-effort communications. *J. Stat. Mech.*, page P11025, 2010.
- [22] Ł. Dębowski. Excess entropy in natural language: present state and perspectives. *Chaos*, 21(3):037105, 2011.